Research Statement

Jonathan H. Huggins

Machine learning and AI are increasingly used for high-impact applications such as law enforcement, pharmacology, and clinical care. But in safety-critical settings, it is crucial for the employed algorithm to be trustworthy. Poor decision-making based on faulty inferences about, for example, the risk of re-offending, the most promising drug candidates, or the pathology of a tumor, could cause real-world harm. Therefore, we need machine learning methods based on reliable algorithms and models that come equipped with rigorous theoretical guarantees. Yet too often such guarantees are unavailable, even for widely-used methods.

In my research, I develop guarantees for probabilistic inference methods and apply these methods to clinical and scientific problems in genomics. In these applications, trusted inferences are important because of the complexity of the data and the health-related implications of inappropriate decisions. My work focuses on probabilistic methods because of benefits such as the ability to incorporate structured knowledge and represent uncertainty. But probabilistic inference poses an algorithmic challenge. For complex models or large amounts of data, computational limitations make it infeasible to perform exact inference. Instead, we must be satisfied with approximate answers. I develop approximate inference algorithms with finite-data, finite-computation guarantees of accuracy. Such guarantees allow a chemist working on drug discovery or a pathologist examining a tumor to be confident that the approximation error will not adversely affect their decision-making.

But even if computation were not a concern, we must still worry about robustness when the model is incorrect—that is, misspecified. Inference could be exact, yet if the model fails to account for important aspects of the data, inferences based on that model and data may still be inaccurate. These concerns have led me to develop methods to assess model robustness that are easy to use and understand.

1 Fast, reliable approximate inference

For many modern statistical problems with large amounts of data or complex models, trustworthy inference has become substantially more difficult because the computational requirements of traditional, trusted methods are too large. For example, Monte Carlo methods, particularly Markov chain Monte Carlo, remain the gold standard for approximate Bayesian inference because of their wide applicability, asymptotic exactness with respect to computation time, strong theoretical guarantees, and reliable convergence diagnostics. However, in many cases it is not feasible to use Markov chain Monte Carlo, so alternative approaches to inference that are more computationally efficient have begun to see wider utilization. For example, popular methods range from the venerable Laplace approximation and its descendant, the integrated nested Laplace approximation to variational methods, which are popular within the machine learning community. Unfortunately, these alternative methods may not be reliable: they lack the finite-data approximation theory and the tools for evaluating approximation accuracy that are available for Markov chain Monte Carlo. The shortcomings of the field's current set of tools raise two possibilities: either create new methods that come equipped with finite-data accuracy guarantees or develop better tools for determining the accuracy of existing approximate inference methods. I have pursued both these avenues in my research.

1.1 Scalable inference via data summarization.

A key insight that I leverage in my work on new algorithms is that in the large-scale setting, much of the data is redundant. Therefore, it is possible to compress data into a form that admits more efficient inference. Approaching approximate inference from this perspective makes computational–statistical tradeoffs conceptually straightforward: less compression translates to greater accuracy but less computational savings.

Coresets. In Huggins et al. (2016), Trevor Campbell, Tamara Broderick and I showed how to summarize a dataset for Bayesian inference by constructing a *coreset*: a small, weighted subset of our data that is representative of the complete dataset. We demonstrated the efficacy of our coreset construction algorithm

on a number of synthetic and real-world datasets. We provided conditions under which the size of the coreset is independent of the original dataset size and found that in practice, the coreset size did not depend on the amount of data. Furthermore, constructing the coreset took a negligible amount of time compared to that required to run Markov chain Monte Carlo. Inspired by these promising preliminary results, my coauthors leveraged the insights we gained in recent follow-up work (Campbell and Broderick, 2017, 2018).

Approximating exponential families. In Huggins et al. (2017a), Ryan Adams, Tamara Broderick, and I compressed the data by first constructing an exponential family that approximates the model of interest and then computing the sufficient statistics of the data. By using (approximately) sufficient statistics, we can summarize the data in a trivially streaming or distributed manner, making only a single pass over the dataset and using a constant amount of memory. With the sufficient statistics in hand, the posterior can then be calculated via, for example, Markov chain Monte Carlo, and point estimates such as the maximum a posteriori solution can be computed—all in time independent of the data set size. We provided rigorous guarantees on the quality of point (maximum a posteriori) estimates, the approximate posterior, and posterior mean and uncertainty estimates. We validated our approach empirically in the case of logistic regression using a quadratic approximation and showed competitive performance with stochastic gradient descent, Markov chain Monte Carlo, and the Laplace approximation in terms of speed and multiple measures of accuracy—including on an advertising data set with 40 million data points and 20,000 covariates. Recently, Zoltowski and Pillow (2018) applied our method to Poisson regression for neural spike train recordings. In addition to empirically showing excellent accuracy on smaller datasets where exact inference was possible, they were able to use our method to a massive real-world dataset with over 2 billion spike count bins.

Theory. An advantage of data compression approaches to approximate inference—including both the coreset and approximating exponential families methods—is that they involve replacing the original likelihood with a (deterministic) approximate likelihood. In Huggins and Zou (2017) and Huggins et al. (2018c), my coauthors and I showed how approximate likelihoods lend themselves to *a priori* analysis of their accuracy. But how do we measure accuracy? To answer this question, we must consider how posterior approximations are used in real-life data analyses. Of central importance is the ability to compute high-quality posterior mean and uncertainty estimates, where uncertainty is typically measured in terms of each parameter's variance or standard deviation. So, to obtain a practically useful finite-data approximation theory, we must carefully consider our choice of divergence measure so as to obtain error bounds on posterior mean and uncertainty estimates. For example, variational inference methods minimize the Kullback–Liebler divergence, but Mikołaj Kasprzak, Trevor Campbell, Tamara Broderick, and I showed in Huggins et al. (2018c), the Kullback–Liebler divergence between two distributions can be small even when the differences between their means and variances is large. The *Wasserstein distance*, on the other hand, provides a better way to determine whether the difference between the exact and approximate posterior is small because closeness in Wasserstein distance between two distributions implies closeness in their means and standard deviations. In Huggins and Zou (2017), James Zou and I developed a framework suited for approximate likelihood methods for bounding the Wasserstein distance between two distributions. We applied our framework to derive finite-sample error bounds of approximate unadjusted Langevin dynamics. In our paper on approximating exponential families (Huggins et al., 2017a), we used the framework from Huggins and Zou (2017) to obtain our finite-data guarantees. In Huggins et al. (2018c), we further developed the tools from Huggins and Zou (2017) and demonstrated their wide applicability: we used them to obtain Wasserstein distance bounds for the approximate posteriors produced by the Bayesian coreset constructions from Campbell and Broderick (2017, 2018) and to the Laplace approximation.

1.2 Scalable nonparametric inference.

Many data analysis problems can be seen as discovering a latent set of traits in a population. For instance, we might recover topics or themes from scientific papers, ancestral populations from genetic data, interest groups from social network data, or unique speakers across audio recordings of many meetings. In all of these cases, we could reasonably expect the number of latent traits present in a data set to grow with the size of the data. However, the compression methods described in Section 1.1 were designed for parametric models, which can only represent a fixed, finite number of traits. Instead, for the types of problems just described, we could

instead use a nonparametric model; that is, a model with an infinite number of parameters, which is able to adapt its complexity to the data. Because they involve an infinite number of parameters, nonparametric models present a distinct set of inferential challenges.

Bayesian nonparametric models. One important class of these models use Bayesian priors based on combinatorial stochastic processes such as the beta and Dirichlet processes. These nonparametric priors provide an infinitely large pool of latent traits so that more and more can be used as the amount of data increases. In Huggins et al. (2017b) and Huggins et al. (2018a), my coauthors and I developed blackbox methods for creating finite approximations to a class of nonparametric priors known as (normalized) completely random measures. This class includes popular priors such as the beta, gamma, and Dirichlet processes. The blackbox methods I developed make it easy for a practitioner to perform posterior inference in these models even when they are not familiar with advanced stochastic process and probability theory. Our approach in Huggins et al. (2018a) was based on the truncation of sequential representations of the prior. In Huggins et al. (2017b, 2018d), Lorenzo Masoero, Lester Mackey, Tamara Broderick, and I used approximations based on having independent and identically distributed parameters, but changing the distribution of the finite parameters as more parameters are included. In this way the finite approximations converge in distribution to the nonparametric prior. We derived error bounds for both types of finite approximations, so that practitioners can easily choose an appropriate approximation level that ensures their results will not differ too much from what they would have obtained with the full nonparametric prior.

Large-scale kernel methods. A very different set of approaches to nonparametric inference are based on using a *kernel function*. A real-valued kernel function takes two data points as arguments and defines a notion of similarity between those observations. By considering the value of the kernel function between every pair of N data points, one can define an $N \times N$ *kernel matrix*. The computational problem with kernel methods is that even computing the kernel matrix requires $O(N^2)$ time and in many cases the kernel matrix needs to be inverted, which in practice requires $O(N^3)$ time. Thus, naïve implementations of kernel methods only scale to a few tens of thousands of observations. As I describe next, I have worked on rigorous approaches to scaling up both Bayesian and non-Bayesian kernel-based methods.

Gaussian processes. *Gaussian processes* (GPs) are a powerful kernel-based method that offer a flexible class of priors for nonparametric Bayesian regression. But popular GP posterior inference approaches are typically prohibitively slow or lack desirable finite-data guarantees on quality. In Huggins et al. (2018b), Mikołaj Kasprzak, Trevor Campbell, Tamara Broderick, and I developed an approach to scalable approximate GP regression with finite-data guarantees on the accuracy of pointwise posterior mean and variance estimates. Our main contribution was a novel objective for approximate inference in the nonparametric setting: the *preconditioned Fisher divergence*. We demonstrated that, for sparse GP likelihood approximations, one can minimize the preconditioned Fisher divergence efficiently. Our experiments showed that optimizing the preconditioned Fisher divergence has the same computational requirements as state-of-the-art scalable GP methods while providing comparable empirical performance—in addition to our novel finite-data quality guarantees.

Linear-time Stein discrepancies. A family of kernels known as *Stein kernels* provides a way to compare distributions when one of the distributions is available via an unnormalized density while the other is only available via sampling. In particular, Stein kernels define kernel Stein discrepancies between distributions, which have been deployed for a variety of applications. For example, they can be used to evaluate the quality of approximate posterior inference algorithms, to perform approximate Bayesian inference, and to do goodness-of-fit testing. Existing convergence-determining kernel Stein discrepancies admit strong theoretical guarantees but suffer from a computational cost that grows quadratically in the sample size because of the need to compute the kernel matrix. While linear-time Stein discrepancies have been proposed for goodness-of-fit testing, they exhibit avoidable degradations in testing power—even when power is explicitly optimized. To address these shortcomings, in Huggins and Mackey (2018), Lester Mackey and I introduced *feature Stein discrepancies* (ΦSDs), a new family of quality measures that can be cheaply approximated using importance sampling. We showed how to construct ΦSDs that provably determine the convergence of a sample to

its target and developed high-accuracy approximations—*random* ΦSD*s* (RΦSDs)—which are computable in near-linear time. In our experiments with sampler selection for approximate posterior inference and goodness-of-fit testing, RΦSDs performed as well or better than quadratic-time kernel Stein distrepancies while being orders of magnitude faster to compute.

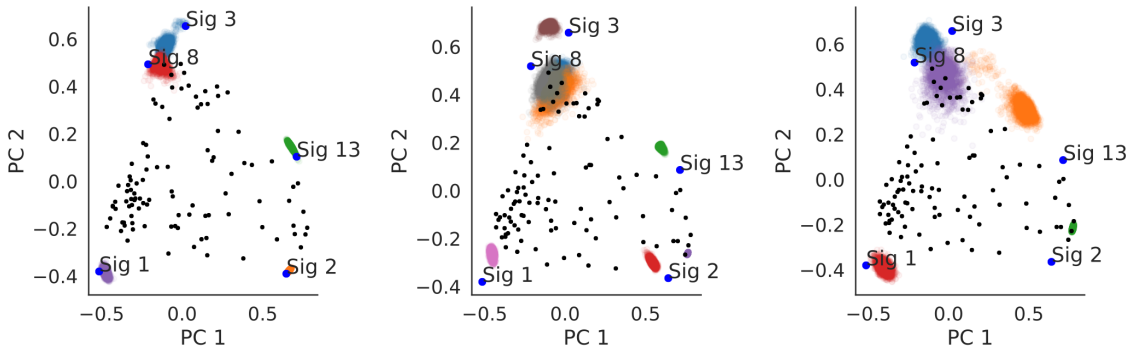
Compressing random features. Two popular approaches for making kernel methods more scalable are random feature maps and the Nyström method, both of which are based on using a low-rank approximation to the kernel matrix. But, in order to achieve desirable theoretical guarantees, the random feature maps may require a prohibitively large number of features J_+ and the Nyström method may be computationally intractable for high-dimensional problems. In Agrawal et al. (2018), we proposed combining the simplicity and generality of random feature maps with a data-dependent feature selection scheme to achieve the desirable theoretical approximation properties of Nyström with just $O(\log J_+)$ features. Our key insight was to begin with a large set of random features, then reduce them to a small number of weighted features in a data-dependent, computationally efficient way, while preserving the statistical guarantees available when using the original large set of features. In addition to our theory showing only $O(\log J_+)$ features are needed to obtain these guarantees, we also demonstrated the efficacy of our method with experiments, including on a data set with over 50 million observations.

2 Inferring biologically relevant structure

When analyzing genome sequencing, gene expression, or other types of noisy high-dimensional data, scientists often use models to discover biologically relevant structure. For example, they may be interested in grouping cells by functionality or understanding how proteins interact to carry out cell processes. However, because of the complexity of the processes that generate the data, model misspecification is a common concern. It is not feasible for the model to accurately represent all aspects of the data generating process both because of complexity and the exact lack of understanding the scientific discovery processes seeks to remedy. When an unsupervised model is misspecified, it often tries to "explain" the data with spurious structure. But using this improperly inferred structure to develop theories or guide future research is fraught, potentially leading to invalid conclusions or wasted resources. Therefore, it is imperative to develop methods that can detect spurious structure or models that are robust against inferring such structures.

Robustly inferring mutational signatures. A revealing example of extraneously inferred structure due to misspecification arises when studying the mutational processes that cause cancer. Tumor cells typically have large numbers of genetic mutations that are *somatic*—that is, mutations that are present in the tumor cells but not in the remaining cells in the body. These somatic mutations are caused by a variety of *mutational processes*, with different cancer types caused by different but overlapping subsets of these processes. Using whole genome and whole exome sequencing of tumor DNA, "signatures" for these mutational processes have been inferred based on unsupervised nonnegative matrix factorization (NMF) methods (Alexandrov et al., 2013). Some mutational signatures have been validated based on biological knowledge; however, because NMF methods are unsupervised and the models are misspecified, there is no "true" set of signatures that can be inferred. Nevertheless, some signatures do seem to correspond genuine mutational processes while others are likely to be spuriously generated due to noise and misspecification. Fig. 1, which is drawn from work with Jeffrey Miller and Scott Carter, demonstrates the drastic effect of even small amounts of misspecification when using Bayesian inference (we see similar results for other inferential approaches). We have developed a method for automatically correcting for misspecification when it is present while also defaulting back to standard Bayesian inference when the data is well-specified. Our robust approach to inference will aid in accurate interpretation of these mutational signatures and in their proper use for downstream applications.

Using mutational signatures for early detection of cancer. One such application of mutational signatures is in my ongoing work with Jeff Miller and Scott Carter's lab, in which we are developing methods for early detection of cancer using blood samples. Cancer detection is possible from a blood draw because human blood contains what is known as *cell-free DNA*, which is DNA from cells that have died. Tumor cells die at a faster rate than is typical and therefore they seem to release more cell-free DNA into the blood



(a) Well-specified (b) Misspecified data (c) Misspecified data with corrected inference

Figure 1: The effects of even minor misspecification can be drastic. The black dots represent the data (used to define to 2D projection using kernel PCA with cosine kernel). The blue dots represent the five mutational signatures used to generate the data. The colorful regions each represent a high-probability region for each inferred signature. In the well-specified case, (a) shows that Bayesian inference correctly infers five signatures and the regions of uncertainty for each signature include a true signature. In the misspecified case, (b) shows that standard Bayesian inference uses many extra, highly uncertain signatures to explain the misspecification. while (c) shows that our corrected method automatically infers the correct number of signatures, although the high-probability regions do not contain all of the true signatures.

stream than non-cancerous cells. Using whole-genome sequencing of cell-free DNA, we were able detect the mutational signatures from cancer when present as well as distinguish between breast cancer, lung cancer, and melanoma. The challenge was that the signal is very weak: we had to carefully filter out apparent mutations that were actually just read errors due to the noisy sequencing process. Our preliminary results are promising: using 29 samples with three cancer types, we were able to correctly classify the type of cancer present 90% of the time.

Inferring stem cell phylogeny from RNA-seq gene expression data. Thanks to recent advances in a technology known as RNA-seq, we can now measure gene expression levels across thousands or millions of individual cells in parallel. Such data has great potential, as it could help us to uncover novel cell types, determine how genes and proteins interact, and learn about how cells differentiate (for example during fetal development or in bone marrow, where blood cells are produced from stem cells). However, this new type of data poses complex computational challenges because it is sparse and noisy, and we must disentangle biological variation from technical variation. In Shiffman et al. (2017, 2018), we have developed a full generative model for probabilistically reconstructing trees of cellular differentiation from single-cell RNA-seq data. Specifically, we extended the framework of the classical nonparametric Dirichlet diffusion tree (DDT) to simultaneously infer branch topology and latent cell state along diffusive trajectories over the full tree. We have demonstrated that Markov chain Monte Carlo inference with our augmented DDT model can recover latent trajectories from simulated single-cell transcriptomes. We are in the process of applying our model to real data.

3 Future work

I have described my research on developing scalable approximate inference algorithms that remain trustworthy and methods for managing model misspecification. But how do computational considerations affect problems of misspecification and vice versa? Below, I highlight three directions for future work where the issues of computation and misspecification intersect.

Frequentist properties of approximate inference algorithms. Analyzing the behavior of approximate inference algorithms in the setting of increasing numbers of observations coming from some fixed distribution can provide invaluable insights into the trade-off between computational complexity and statistical accuracy. I am particularly interested in further studying the large-data behavior of the coreset and approximate exponential family methods I helped develop. Understanding the inherent computational–statistical tradeoffs of these methods can guide users in deciding when a method is likely to perform well.

Model misspecification and computation. How do we manage the widespread issue of misspecification? While there are many available approaches to dealing with misspecified models, I am interested in developing more automated and computationally efficient methods to managing misspecification that will be easy for practitioners to use. Providing these easier to use tools will lead to an increased ability to manage misspecification when confronted with it. I would also like to help deepen our understanding of the computational benefits or problems that may arise from misspecification. For example, forcing the posterior to be less concentrated in order to combat overconfidence could lead to a more diffuse posterior that is easier to approximate.

Robust, large-scale exploratory data analysis of cancer genomes and transcriptomes. As I outlined in Section 2, exploratory data analysis is an important step in the scientific discovery process, helping to generate promising directions for future research. However, when applied to large amounts of noisy, high-dimensional data now available, standard approaches to clustering, non-negative matrix factorization, manifold learning, and other unsupervised problems may either be computationally infeasible or produce unreliable results that are not robust to noise or spurious correlations. Therefore, an important direction for my research is to develop more robust, scalable approaches to exploratory data analysis. My work in mutational signature inference is an example of one such direction I have already pursued. I am also interested in the problem of finding manifold structure in RNA-seq data, which is a low-dimensional representation of the phenotypic space cells reside in. Existing methods such as diffusion maps do not scale to the 100,000+ cells that can now be sequenced and are unable to properly account for the large amounts of observation noise. One promising approach is to use coreset-based inference methods for Bayesian manifold learning models, which would be scalable while also being robust to the low signal-to-noise data.

References

Raj Agrawal, Trevor Campbell, **Jonathan H. Huggins**, and Tamara Broderick. Data-dependent compression of random features for large-scale kernel approximation. *arXiv.org*, 2018.

Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Samuel A J R Aparicio, Sam Behjati, et al. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, Aug. 2013.

Trevor Campbell and T. Broderick. Automated Scalable Bayesian Inference via Hilbert Coresets. *arXiv.org*, stat.ML:1710.05053, Oct. 2017.

Trevor Campbell and T. Broderick. Bayesian Coreset Construction via Greedy Iterative Geodesic Ascent. In *International Conference on Machine Learning*, pages 1–13, 2018.

Jonathan H. Huggins and Lester Mackey. Random feature Stein discrepancies. In *Proc. of the 32nd Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

Jonathan H. Huggins and James Zou. Quantifying the accuracy of approximate diffusions and Markov chains. *Proc. of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, PMLR 54:382–391, 2017.

Jonathan H. Huggins, Trevor Campbell, and Tamara Broderick. Coresets for scalable Bayesian logistic regression. *Proc. of the 30th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2016.

Jonathan H. Huggins, Ryan P. Adams, and Tamara Broderick. PASS-GLM: polynomial approximate sufficient statistics for scalable Bayesian GLM inference. *Proc. of the 31st Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2017a.

Jonathan H. Huggins, Lorenzo Masoero, Lester Mackey, and Tamara Broderick. Generic finite approximations for practical bayesian nonparametrics. In *NeurIPS Workshop on Advances in Approximate Bayesian Inference*, 2017b.

Jonathan H. Huggins, Trevor Campbell, Jonathan P. How, and Tamara Broderick. Truncated random measures. *Bernoulli*, 2018a.

Jonathan H. Huggins, Trevor Campbell, Mikołaj Kasprzak, and Tamara Broderick. Scalable Gaussian process inference with finite-data mean and variance guarantees. *arXiv.org*, 2018b.

Jonathan H. Huggins, Mikołaj Kasprzak, Trevor Campbell, and Tamara Broderick. Bayesian posterior mean and uncertainty estimates: a non-asymptotic approach. *arXiv.org*, 2018c.

Jonathan H. Huggins, Lorenzo Masoero, Lester Mackey, and Tamara Broderick. Non-nested finite approximations for bayesian nonparametric priors. *In preparation*, 2018d.

Miriam Shiffman, Will Stephenson, Geoffrey Schiebinger, Trevor Campbell, **Jonathan H. Huggins**, Aviv Regev, and Tamara Broderick. Probabilistic reconstruction of cellular differentiation trees from single-cell RNA-seq data. In *NeurIPS Workshop on Machine Learning in Computational Biology*, 2017.

Miriam Shiffman, Will Stephenson, Geoffrey Schiebinger, **Jonathan H. Huggins**, Trevor Campbell, Aviv Regev, and Tamara Broderick. Reconstructing probabilistic trees of cellular differentiation from single-cell RNA-seq data. *arXiv.org*, 2018.

David Zoltowski and J. W. Pillow. Scaling the Poisson GLM to massive neural datasets through polynomial approximations. In *Advances in Neural Information Processing Systems*, 2018.